基于多粒度粗糙集的聚类融合方法 *

于佩秋 [1,2]，李进金 [1†]，林国平 [1,2]

(1. 闽南师范大学 数学与统计学院，福建 漳州 363000; 2. 福建省粒计算重点实验室，福建 漳州 363000)

摘　要：现有的聚类融合算法从聚类成员的角度出发，若使用全部聚类成员则融合结果受劣质成员影响，对聚类成员进行选择再进行融合则选择的策略存在主观性。为在一定程度上避免这两种局限性，可以从元素的角度出发，提出一种新的聚类融合方法。通过多粒度决策不一致粗糙集来选择一部分类别确定的元素，再利用这部分元素进行聚类融合生成新的划分；多粒度决策不一致粗糙集模型能够刻画多粒度决策过程中属性一致而决策不一致的现象，提出了一种基于多粒度决策不一致的粗糙集模型，并给出了一种聚类融合方法。具体做法是：首先在数据集上多次使用 K-means 聚类算法，生成论域上的多个粒结构；其次对所有粒结构两两之间求粒间包含度，建立包含度矩阵，对矩阵使用 Otsu 算法计算阈值，得出多组满足阈值条件的信息粒，求解多粒度决策不一致下近似和上近似；最后分别处理下近似与边界域中元素的类别，从而获得了一个经过融合的聚类划分。实验结果表明，该方法能够有效改善聚类的结果，具有较高的时间效率，且算法具有较好的鲁棒性。

关键词：多粒度粗糙集；聚类融合；大津算法；包含度

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.04.0217

Clustering ensemble algorithm based on multi-granulation rough set

Yu Peiqiu[1,2], Li Jinjin[1†], Lin Guoping[1,2]

(1. *School of Mathematics & Statistics*, *Minnan Normal University*, *Zhangzhou Fujian* 363000, *China*; 2. *Laboratory of Granular Computing*, *Zhangzhou Fujian* 363000, *China*)

Abstract: Existing clustering ensemble algorithm starts from the perspective of cluster members, if all the cluster members are used, the ensemble result is affected by the inferior members. If the cluster members are selected and then used in ensemble, the selected strategy has subjectivity. To avoid these two limitations to some extent, from the perspective of elements, nature proposes a new clustering fusion method: selecting a part of class-determined elements through multi-granulation rough sets with incongruous decisions, and then using this part of the elements to generate a new clustering. Multi-granulation rough set model with incongruous decisions can describe the phenomenon of inconsistent decisions with consistent attribute set, a model of multi-granulation rough set with incongruous decisions and a clustering ensemble algorithm based on the model were proposed in this paper. First of all, run a K-Means clustering algorithm several times on the data set in the case, multiple granule structures were generated. Next, inclusion degrees among all the granulations were calculated, and then the matrix of inclusion degree was obtained. Used Otsu's method to generate a threshold, then several group of granulation that met the threshold condition were got. According to the model of multi-granulation rough set with incongruous decision, lower and upper approximations were obtained. Finally, classified the elements of lower approximation and boundary separately, then a clustering that has been fused was obtained. The experiments showed that the algorithm had a high time efficiency and robustness, which improved the result of K-means clustering.

Key words: multi-granulation rough set; clustering ensemble; Otsu's method; inclusion degree

0　引言

聚类分析[3]是在探索性数据分析领域尤其是在数据挖掘和知识发现方面的一种重要方法,用以揭示数据分布的真实情况。聚类分析目前已被成功应用于工程、生物学、心理学、药学等其他学科中.目前已有的聚类算法还不能够胜任对任意分布情况以及任意形状的数据的聚类,传统的聚类算法都是为特定领域而设计的,在伸缩性和稳定性等方面存在种种不足,因此引入聚类融合[4,5,6,15,17],对聚类结果进行合并,从而得到比单次运行聚类算法更为优越的结果.聚类融合是一个非常强大的工具,可以大大提高非监督分类方法的健壮性以及稳定性。经典的多粒度粗糙集模型[1,2]以属性集上的子集来确定不同的划分,从而形成多个粒度,没有考虑属性集完全相同而决策不同的情况。本文提出了一种刻画属性集相同而决策不同的现象的多粒度决策不一致粗糙集模型,丰富和发展了多粒度粗糙集理论。在使用聚类算法生成划分的过程中,经常存在聚类算法给出的类别标签不一致,这种情况是多粒度决策不一致粗糙集模型的一个特例,可以在聚类融合时使用多粒度决策不一致粗糙集模型。聚类融合是

收稿日期：2018-04-03；**修回日期：**2018-05-23　**基金项目：**福建省自然科学基金资助项目（2016J01315，2017J01507）；国家自然科学基金资助项目（61379021）；国家青年科学基金资助项目（61603173）；浙江省海洋大数据挖掘与应用重点实验室开放课题（OBDMA201603）；2017 年福建省中青年教师教育科研项目（JAT170340）；福建省数学类研究生教育创新基地资助项目（1013-313009）

作者简介：于佩秋（1991-），男，硕士研究生，主要研究方向为粗糙集、粒计算、数据挖掘、人工智能；李进金（1960-），男（通信作者），教授，博士，主要研究方向为人工智能、粒计算、拓扑学（jinjinli@mnnu.edu.cn）；林国平（1978-），副教授，博士，主要研究方向为多粒度粗糙集、信息融合、粒计算、数据挖掘、人工智能.

基于聚类分析的结果而产生的一种融合策略。

对于聚类融合,目前一部分学者采用对已有的所有聚类结果进行进一步融合的分析逻辑[18]，例如李飞江等人[5]提出了一种粗糙集和证据理论相结合的聚类融合方法，Fred[7]利用数据点之间的相似度建立共生矩阵，通过设置阈值来判断矩阵中的两个点是否属于聚类结果中的同一类，此外还有 Srehl 和 Ghosh[8]提出了三个基于超图的方法 MCLA，HGPA 和 CSPA.这些方法都是对所有的聚类结果进行融合,不能避免劣质聚类成员对聚类融合的质量产生的影响.另一部分学者首先对聚类成员进行评价[16],剔除劣质聚类成员而后再进行聚类融合,例如 Faceli 等人[9]通过遗传算法迭代优化得到最优的融合结果；Hong 等人[10]提出通过首先对聚类成员进行选择来提高最终聚类融合结果的质量；阳琳赟等人提出了一种基于粗糙集理论的聚类融合加权迭代模型。这些方法对聚类成员进行了选择,但聚类成员的评价和选择具有较强的主观性,从而使聚类融合结果产生一定程度上偏差。使用多粒度决策不一致粗糙集模型求解真实聚类的下近似，由下近似中元素的类别决定边界域中元素的类别的方法可以在一定程度上减弱这种由于聚类成员的选择而产生的偏差。

事实上无论聚类成员优劣程度如何,对同一真实聚类而言,劣质聚类成员与优质聚类成员对某些元素的归属可以达成共识。由于多粒度粗糙集具有求同存异的特点[1]，本文首次尝试借鉴多粒度粗糙集理论求多粒度下近似的方法求劣质聚类成员与优质聚类成员的共识元素。将在一个完备的信息系统中多次运行 K-means 聚类算法生成多个划分,即多个粒度.每个划分中的聚类成员视为等价类,利用多粒度融合的方法,求这些聚类成员的下近似和上近似。通过考量下近似中元素和边界域中元素之间的关系,利用"类间差异大,类内差异小"的聚类基本原则,通过求距离所有下近似距离最近的一个元素与每个下近似中部分紧邻元素的平均距离的最小值来决定该元素的分类,从而可以重新建立划分。

为了进行后续的讨论,首先介绍多粒度决策不一致粗糙集模型。

1　多粒度决策不一致粗糙集

在客观世界的决策过程中,由于决策是由专家给出的,存在主观性,即基于同样的条件,不同专家给出的决策可能会不同,这种现象在本文中被称为多粒度决策不一致。首先给出多粒度决策不一致信息系统的概念。

定义 1　多粒度决策不一致信息系统。设信息系统$MS = \{IS_i | IS_i = (U, AT, f_i)\}(i \le m)$为多粒度信息系统，其中$IS_i = (U, AT, f_i)$为一个三元信息系统，$U = \{x_1, x_2, \cdots, x_n\}$为非空有限论域；$AT = \{a_1, a_2, \cdots, a_{|AT|}\}$为属性集；$f_i$：$U \times AT \longrightarrow V_c$为决策函数,$V_c$为决策指标集，即$\forall x \in U$有$f(x, AT) \in V_c$.若$\exists x \in U$,当$1 \le r \le m, 1 \le s \le m$,使得$f_r(x) \ne f_s(x)$，称$MIDS = \{IS_i | IS_i = (U, AT, f_i)\}(i \le m)$为多粒度决策不一致信息系统。

然后借鉴多粒度粗糙集思想[1],定义多粒度决策不一致粗糙集模型。

定义 2　多粒度决策不一致粗糙集。设$MIDS = \{IS_i | IS_i = (U, AT, f_i), i = 1,2,\cdots,m\}$是一个多粒度决策不一致信息系统，$IS_i = (U, AT, f_i), f_i$：$U \times AT \longrightarrow V_c$为决策函数，则多粒度决策不一致下近似为

$$\underline{ID_{\sum_{i=1}^m f_i}}(x) = \{(y \in U | f_1(x) = f_1(y) \wedge f_2(x) = f_2(y) \wedge \cdots \wedge f_m(x),$$

多粒度决策不一致上近似为

$$\overline{ID_{\sum_{i=1}^m f_i}}(x) = \{y \in U | f_1(x) = f_1(y) \vee f_2(x) = f_2(y) \vee \cdots \vee f_m(x),$$

多粒度决策不一致边界为

$$BN_{\sum_{i=1}^m f_i}(x) = \overline{ID_{\sum_{i=1}^m f_i}}(x) - \underline{ID_{\sum_{i=1}^m f_i}}(x);$$

那么称$\left(\underline{ID_{\sum_{i=1}^m f_i}}(x), \overline{ID_{\sum_{i=1}^m f_i}}(x)\right)$为多粒度决策不一致粗糙集模型。

多粒度决策不一致粗糙集具有如下性质：

(1) $\underline{ID_{\sum_{i=1}^m f_i}}(x) \subset \overline{ID_{\sum_{i=1}^m f_i}}(x)$;

(2) $\cup_{x \in U} \underline{ID_{\sum_{i=1}^m f_i}}(x) = U, \cup_{x \in U} \overline{ID_{\sum_{i=1}^m f_i}}(x) = U$;

(3) $\forall u \in \underline{ID_{\sum_{i=1}^m f_i}}(x), \underline{ID_{\sum_{i=1}^m f_i}}(u) = \underline{ID_{\sum_{i=1}^m f_i}}(x)$;

(4) $\underline{ID_{\sum_{i=1}^m f_i}}(x) = \cap_{i=1}^m \underline{f_i}(x), \overline{ID_{\sum_{i=1}^m f_i}}(x) = \cup_{i=1}^m \overline{f_i}(x)$;

证明：

(1) $\forall y \in \underline{ID_{\sum_{i=1}^m f_i}}(x)$ 有 $f_i(y) = f_i(x)(\forall i \le m) \Rightarrow y \in \overline{ID_{\sum_{i=1}^m f_i}}(x), \Rightarrow \underline{ID_{\sum_{i=1}^m f_i}}(x) \subset \overline{ID_{\sum_{i=1}^m f_i}}(x)$.

(2) $\forall x \in U$有$f_i(x) = f_i(x)(\forall i \le m) \Rightarrow U \subseteq \cup_{x \in U} \underline{ID_{\sum_{i=1}^m f_i}}(x)$，另一方面 $\forall x \in U, \underline{ID_{\sum_{i=1}^m f_i}}(x) \subseteq U \Rightarrow \cup_{x \in U} \underline{ID_{\sum_{i=1}^m f_i}}(x) = U$,类似可证$\cup_{x \in U} \overline{ID_{\sum_{i=1}^m f_i}}(x) = U$.

(3) $\forall u \in U, u \in \underline{ID_{\sum_{i=1}^m f_i}}(x) \Leftrightarrow \forall i \le m, f_i(u) = f_i(x) \Leftrightarrow \underline{ID_{\sum_{i=1}^m f_i}}(u) = \underline{ID_{\sum_{i=1}^m f_i}}(x)$.

(4) 只证$\underline{ID_{\sum_{i=1}^m f_i}}(x) = \cap_{i=1}^m \underline{f_i}(x)$，后面的部分类似可证；$\forall x \in \underline{ID_{\sum_{i=1}^m f_i}}(x)$,满足$f_1(x) = f_1(y) \wedge f_2(x) = f_2(y) \wedge \cdots \wedge f_m(x)$，则$x \in \cap_{i=1}^m \underline{f_i}(x)$，另一方面 $\forall x \in \cap_{i=1}^m \underline{f_i}(x)$ 有$f_1(x) = f_1(y) \wedge f_2(x) = f_2(y) \wedge \cdots \wedge f_m(x)$, 即 $x \in \underline{ID_{\sum_{i=1}^m f_i}}(x)$，综上所述$\underline{ID_{\sum_{i=1}^m f_i}}(x) = \cap_{i=1}^m \underline{f_i}(x)$.

例 1　多粒度决策不一致信息系统与上下近似的求解

设论域$U = \{x_1, x_2, x_3, x_4, x_5\}$,在两个不同的粒度下分别有如表 1 所示的决策表。

表 1　一个多粒度决策不一致信息系统

U	A	D_1	U	A	D_2
x_1	1	1	x_1	1	1
x_2	1	1	x_2	1	1
x_3	2	2	x_3	2	2
x_4	2	2	x_4	2	2
x_5	3	2	x_5	3	1

显然$MIDS = \{IS_i|IS_i = (U, A, f_i), i = 1,2\}$是一个多粒度决策不一致信息系统.由定义 1.2 有：$\underline{ID_{\sum_{i=1}^m f_i}}(x_1) = \underline{ID_{\sum_{i=1}^m f_i}}(x_2) = \{x_1, x_2\}$，$\overline{ID_{\sum_{i=1}^m f_i}}(x_1) = \overline{ID_{\sum_{i=1}^m f_i}}(x_2) = \{x_1, x_2, x_5\}$，$BN_{\sum_{i=1}^m f_i}(x_1) = BN_{\sum_{i=1}^m f_i}(x_2) = \{x_5\}$.

2 基于多粒度粗糙集的聚类融合算法（MGIDA）的设计

2.1 噪声的去除和多粒度不一致下近似与边界域的求解

首先介绍一种本文所用的聚类算法：K-means 聚类算法：

K-means 聚类算法是一种至今仍然广泛应用的经典聚类算法，该算法流程如下：设聚类类别数目为k，

1.选定k个初始聚类中心$K = \{\{x_1\}, \{x_2\}, \cdots, \{x_k\}\}$，分别代表$k$个类，此时$K_1 = \{x_1\}, K_2 = \{x_2\}, \cdots, K_k = \{x_k\}$;

2.重复以下计算过程，直到所有的聚类中心不再改变：

(1)对每个$x \in U - K$,计算x到k个聚类中心的距离，假设x到K_*的聚类中心的的距离是最近的则使$K_* = K_* \cup x$;

(2) 重新计算K_*内样本的各项属性的平均值作为K_*的新的聚类中心.

3. 输出所有的聚类中心和元素类别，算法结束.

设U为非空有限论域，在该论域上多次运行聚类算法后生成论域多个 U 的划分，把每次运行聚类算法所形成的划分看做是单个粒度结构,多次运行聚类算法形成多个粒度结构,即多个粒度空间。利用定义 2 的性质（4）可以方便地求得该粒度空间中多粒度决策不一致下近似。但在求解的过程中，可能存在由一致性噪声产生的下近似，在利用多粒度不一致粗糙集模型进行融合的之前，需要去除这些噪声。方法是首先对论域上的所有粒求粒间包含度。

定义 3[11] 设集合A与B是论域U的非空子集,定义集合A与集合B的包含度为

$$sim(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

并将计算得到的包含度存入包含度矩阵$S(C)$,包含度矩阵定义如下:

定义 4 包含度矩阵$S(C)$。设有论域U上的一个子集族$C = \{C_n: C_n \subseteq U\}$,由定义 2.1 计算$C_i(i < n)$与$C_j(j < n)$之间的相容度,并将$C_i$与$C_j$的相容度填入矩阵$S(C)$的第$i$行第$j$列所获得的矩阵即为包含度矩阵。

显然有任意的$S(C)$中的值大于等于 0 且小于等于 1。得到包含度矩阵后,利用 Otsu 算法[12]对该矩阵计算包含度阈值。Otsu 算法又称大津算法，是由日本学者大津于 1979 年提出的一种使前景与背景的类间方差最大化的阈值方法.用以对图像进行阈值分割,该方法又称为最大类间方差法.设$S(C)_{r\times r}$的均值为m,存在阈值t将S(C)中的所有元素分为两类，大于t的类A和小于t的类B, A的均值为m_A, B的均值为m_B, Nobuyuki Otsu[12]给出的类间方差定义为

$$I_{AB} = \frac{|A|}{r^2}(m_A - m)^2 + \frac{|B|}{r^2}(m_B - m)^2;$$

该方法寻找一个最佳阈值$\hat{t}$使得将用灰度值表示的图像分割为两类后错分概率最小,这个最佳的阈值即遍历t的各种取值，取使得I_{AB}最大的一个 t

本文将包含度矩阵$S(C)$作为图像，计算得到阈值后,小于阈值的包含度即由一致性噪声造成的(可视为背景)。对包含度大于阈值的粒求多粒度决策不一致下近似即可。由定义 2 的性质 (4), 设满足阈值条件的信息粒为$\{C_{s_1}, C_{s_2}\}$, 则$\forall x \in \cup_{i=1}^2 C_{s_i}$, $\underline{ID_{\sum_{i=1}^m f_i}}(x) = \cap_{i=1}^2 C_{s_i}$. 若存在$\underline{ID_{\sum_{i=1}^m f_i}}(x) \cap \underline{ID_{\sum_{i=1}^m f_i}}(y) \neq \emptyset$，则合并这两个下近似，使$\underline{ID_{\sum_{i=1}^m f_i}}(x) = \underline{ID_{\sum_{i=1}^m f_i}}(y) = \underline{ID_{\sum_{i=1}^m f_i}}(x) \cup \underline{ID_{\sum_{i=1}^m f_i}}(y)$。在以后的讨论中，令$BN = U - \cup_{x\in U} \underline{ID_{\sum_{i=1}^2 f_i}}(x)$为边界域。

例 2 去除噪声和求解多粒度决策不一致下近似与边界域

设论域$U = \{x_1, x_2, x_3, x_4, x_5\}$,使用聚类算法运行两次生成的划分为$C = \{C_1, C_2\}$,其中$C_1 = \{\{x_1, x_2, x_5\}, \{x_3, x_4\}\}$, $C_2 = \{\{x_1, x_2\}, \{x_3, x_4, x_5\}\}$,计算相容度并填入相容度矩阵，得到相容度矩阵$S(C)$:

$$S(C) = \begin{bmatrix} 1 & 0 & 2/3 & 1/5 \\ 0 & 1 & 0 & 2/3 \\ 2/3 & 0 & 1 & 0 \\ 1/5 & 2/3 & 0 & 1 \end{bmatrix}$$

使用 Otsu 算法计算阈值(在 MATLAB 数学软件中，Otsu 算法是一个系统函数 graythresh),得到阈值为 0.4314,则满足阈值条件的信息粒为$Gr_1 = \{\{x_1, x_2, x_5\}, \{x_1, x_2\}\}$，$Gr_2 = \{\{x_3, x_4\}, \{x_3, x_4, x_5\}\}$,不满足阈值条件的包含度即由一致性噪声造成，不作处理。求对满足阈值条件的信息粒解多粒度决策不一致下近似并求边界域,得

$$\underline{ID_{\sum_{i=1}^m f_i}}(x_1) = \underline{ID_{\sum_{i=1}^m f_i}}(x_2) = \{x_1, x_2, x_5\} \cap \{x_1, x_2\} = \{x_1, x_2\};$$

$$\underline{ID_{\sum_{i=1}^m f_i}}(x_3) = \underline{ID_{\sum_{i=1}^m f_i}}(x_4) = \{x_3, x_4\} \cap \{x_3, x_4, x_5\} = \{x_3, x_4\};$$

$$BN = U - \underline{ID_{\sum_{i=1}^m f_i}}(x_1) \cup \underline{ID_{\sum_{i=1}^m f_i}}(x_3) = \{x_5\}.$$

由定义 2，下近似中的元素在每次聚类的过程中都同属于一类，所以下近似中元素是确定同属一类的,而边界中的元素则不一定在每次聚类过程中都同属于一类，所以边界域中元素类别是不定的,所以需要设计算法确定边界域中元素类别。

2.2 边界域元素的处理

为了方便处理边界域的元素，首先给出聚类的一种定义，该定义是根据“类间距离大，类内距离小”得出的。

定义 5 设(U, A, f)为完备信息系统，给定距离度量

$d(x,y): U\times U\rightarrow[0,+\infty)$，聚类即在$U$上建立划分C= $\{C_k: k=1,2,\cdots,m\}$，使得对于任意的$x,y\in C_i$，若满足对于任意的$y'\in C_i, d(x,y)\le d(x,y')$，则对于任意的$z\in C_j(i\ne j)$有$d(x,y)<d(x,z)$.

聚类融合最终是要生成一个聚类.使用多粒度决策不一致粗糙集模型不可避免地产生边界域,如何处理这些边界域中元素将是一个值得讨论的问题;如果能够找到多粒度决策不一致下近似中元素与边界域中元素的某种关系,即可通过下近似中的元素的类别来确定边界域中元素的类别。由定义 5 给出如下定理:

定理 1 设(U,A,F)为完备信息系统，C= $\{C_k: k=1,2,\cdots,m\}$是U上的一个聚类，$\underline{C}=\left\{\underline{C_k}: \underline{C_k}\subseteq C_k, k=1,2,\cdots,m\right\}$，其中$\underline{C_k}$是$C_k$的任意非空子集，$x\in U-\cup_{i=1}^{m}\underline{C_i}$且$d(x,y)=min\left\{D\left(x,\underline{C_k}\right): k=1,2,\cdots,m\right\}$对任意$y\in\cup_{k=1}^{m}\underline{C_k}$都成立，则$x\in C_i$当且仅当存在$y\in\underline{C_i}$使得对于任意的$z\in\underline{C_j}(i\ne j)$有$d(x,y)<d(x,z)$，其中$D(x,X)=min\{d(x,t): t\in X\}$。

证明：1）充分性

若$d(x,y)=min\left\{D\left(x,\underline{C_k}\right): k\le m\right\}$且$y\in\cup_{k=1}^{m}\underline{C_k}$，则

$\exists\underline{C_i}, s.t. d(x,y)=D\left(x,\underline{C_i}\right), \underline{C_i}\subseteq C_i\Rightarrow\forall z\in U-\underline{C_i}, d(x,y)<d(x,z)\Rightarrow\forall z\in U-C_i\subseteq U-\underline{C_i}, d(x,y)<d(x,z)\Rightarrow x\in C_i$。

2）必要性

若$x\in C_i, x\in U-\cup_{k=1}^{m}\underline{C_k}$且$d(x,y)=min\left\{D\left(x,\underline{C_k}\right): k=1,2,\cdots,m\right\}$对任意$y\in\cup_{k=1}^{m}\underline{C_k}$都成立，则$y\in\underline{C_i}$且$\forall z\in\underline{C_j}(i\ne j)$有$d(x,y)<d(x,z)$.否则$y\in\underline{C_j}(i\ne j), \exists t\in\underline{C_i}$且满足$d(x,t)=\min\{d(x,r)|r\in\underline{C_i}\}$使$d(x,y)<d(x,t)\Rightarrow x\in C_j$，矛盾。

说明:定理 1 给出了一种处理边界域元素的方式，即距离所有下近似最近的一个元素（最小的$D\left(x,\underline{C_k}\right)$，$k\le m$）一定属于距离这个元素最近的一个下近似。可以通过寻找距离所有下近似最近的一个边界域元素x并将该元素并入距离它最近的一个下近似以逐步缩小边界域。由于真实数据集存在复杂性，所以比较可靠的方式是用x到下近似中最近的一部分元素（元素个数为N_0）的平均距离来代替x到下近似中最近的一个元素的距离.然后通过比较x到所有下近似距离的最小值即可得出边界域中每个元素的归属。在本文中取$N_0=\frac{min\left\{\left|\underline{C_k}\right| | k=1,2,\cdots,m\right\}}{2}+1$。重复这一过程，当边界域中所有的元素都被并入下近似后，论域中不再存在类别不确定的元素，即形成了一个新的划分。

例 3 边界域元素的处理

续例 2，$BN=U-\underline{ID_{\sum_{i=1}^{m}f_i}}(x_1)\cup\underline{ID_{\sum_{i=1}^{m}f_i}}(x_3)=x_5, BN=BN-\{x_5\}, N_0=\frac{\min\{2,2\}}{2}+1=2$，设$\underline{Gr_1}$代表由$Gr_1$求得的下近似，$\underline{Gr_2}$代表由$Gr_2$求得的下近似，若$\frac{d(x_5,x_1)+d(x_5,x_2)}{2}<\frac{d(x_5,x_3)+d(x_5,x_4)}{2}$，则$\underline{Gr_1}=\underline{Gr_1}\cup\{x_5\}$，否则$\underline{Gr_2}=\underline{Gr_2}\cup\{x_5\}$。

基于以上讨论,给出基于多粒度决策不一致粗糙集的聚类融合算法：

算法 1 基于多粒度决策不一致粗糙集的聚类融合算法(MGIDA)

输入：运行多次或多个聚类算法生成的一族划分。

输出：一个经过聚类融合的划分。

Step1 计算相容度矩阵；

Step2 使用 Otsu 算法计算相容度矩阵的阈值，聚类成员之间相容度大于阈值即一个满足阈值条件的信息粒，求出所有这样的信息粒；

Step3 利用定义 2 求解下近似，求边界域BN；

Step4 取$x\in BN$且满足$d(x,y)=min\left\{D\left(x,\underline{ID_{\sum_{i=1}^{m}f_i}}(x)\right): k=1,2,\cdots,m; x\in U\right\}$，使用定理 1 的说明的方法对$x$重新归类；

Step5 $BN\leftarrow BN-\{x\}$；当$BN\ne\emptyset$时转至 Step4；

Step6，输出所有元素的类别。

为了验证算法的有效性.下面在 10 个数据集上进行验证。

3 实验结果对比与分析

使用的数据集相关的信息如表 3 所示。为使同一数据集产生不同的较差的划分,使用算法 2 对数据集进行处理,算法的具体过程如例 4。

例 4 使用同一数据集生成不同的弱划分

给定信息系统如表 2 所示，分别生成 2 个模为 1 的一维随机向量：

$$r_1=\langle 0.5030,0.8406,0.2007\rangle^T,$$

$$r_2=\langle 0.0979,0.6985,0.7089\rangle^T;$$

分别使$U\cdot r_1, U\cdot r_2$其中‘·’代表矩阵乘法,得到：

$$U_{r_1}=\langle 0.6967,0.4586,0.8946,1.3101\rangle^T,$$

$$U_{r_2}=\langle 0.6454,0.4254,0.8774,0.9974\rangle^T.$$

表 2 一个不带决策的信息系统

U	a_1	a_2	a_3
x_1	0.4173	0.4929	0.3692
x_2	0.0497	0.4893	0.1112
x_3	0.9027	0.3377	0.7803
x_4	0.9448	0.9001	0.3897

分别对U_{r_1}、U_{r_2}使用 K-means 聚类算法进行聚类，得到两个不同的弱划分：

$$C_1 = \{\{x_1, x_2, x_3\}, \{x_4\}\}, C_2 = \{\{x_1, x_2\}, \{x_3, x_4\}\},$$

这样就使用同一数据集生成了不同的弱划分。

算法 2 [5]弱划分的生成

1：生成一个随机的d维随机向量u,并使$|u| = 1$

2：$X' = X_{n\times d} \cdot u_{d\times 1}$

3：$A_m \leftarrow KMeans(X')(m < n)$

使用 K-means 算法对处理后的数据集进行聚类.使用 CSPA、HGPA、MCLA（前三者均为基于图的聚类融合算法）、IWCE[13]（基于粗糙集理论的聚类融合加权迭代模型）、DSCE[5]（多粒度信息融合：一种基于证据理论的聚类集成方法）作为对比算法。

表 3 实验使用的 UCI 数据集

ID	数据集	实例数	属性	类数
1	Turkey Student Evaluation Generic	5820	31	13
2	Epileptic Seizure Recognition Data Set	11500	178	5
3	Data User Modeling Dataset	258	5	4
4	Synthetic control Data	600	60	6
5	Seeds data set	210	7	3
6	Wine Recognition data	178	13	3
7	Iris	150	4	3
8	Mammographic Mass Data	830	5	2
9	Pima Indians Diabetes Database	768	8	2
10	HTRU2	17898	8	2

每次生成 4 个弱划分进行聚类融合,融合结果与真实的聚类对比计算聚类精度[11],聚类精度定义如下：

$$AC = \sum_{i=1}^{k} \frac{max_{j=1,2,\cdots,k} n_{ij}}{n};$$

其中,若真实的聚类划分为$C_R = \{C_1, C_2, \cdots, C_k\}$,聚类融合得到的聚类划分为$C_F = \{F_1, F_2, \cdots, F_k\}$,则

$$n_{ij} = |C_i \cap F_j|, i, j \le k;$$

表 4 UCI 数据对比实验结果

	CSPA	HGPA	MCLA	IWCE	DSCE	MGIDA
1	0.1739±	**0.1742**	0.1713	**0.1742**	0.1688	0.1683
	0.0000	**±0.0000**	±0.0000	**±0.0000**	±0.0000	±0.0000
2	**0.2729±**	0.2097	0.2385	0.1740	0.1284	0.2477
	0.0000	±0.0000	±0.0001	±0.0000	±0.0000	±0.0000
3	0.4744	0.4074	0.4589	0.4780	0.4613	**0.4992**
	±.0011	±.0009	±0.0021	±0.0031	±0.0024	**±0.0152**
4	0.6370±	0.5195	0.6518	**0.6573**	0.5676	0.6460
	0.0056	±0.0068	±0.0002	**±0.0007**	±0.3200	±0.0009
5	0.7652±	0.5790	0.7300	**0.7784**	0.6730	0.7329
	0.0054	±0.0097	±0.0071	**±0.0034**	±0.0049	±0.0092
6	0.7213±	0.6725	0.7803	0.725	0.7743	**0.7989**
	0.0005	±0.0022	±0.0077	0±0.0007	±0.0112	**±0.0152**
7	0.8810±	0.4673	**0.9200±**	0.8714	0.8079	0.8787
	0.0022	±0.0098	**0.0012**	±0.0015	±0.0174	±0.0180
8	0.8560	0.6035	**0.9165**	0.8544	0.8684	0.8897
	±0.0002	±0.0000	**±0.014**	±0.0008	±0.0139	±0.0081
9	0.7852±	0.7852	0.9018	0.7852	**0.9192**	0.9053
	0.0000	±0.0000	±0.0006	±0.0000	**±0.0028**	±0.0068
10	0.9084	0.9084	0.9084	0.9084	0.9084	**0.9087**
	±0.0000	±0.0000	±0.0000	±0.0000	±0.0000	**±0.0000**

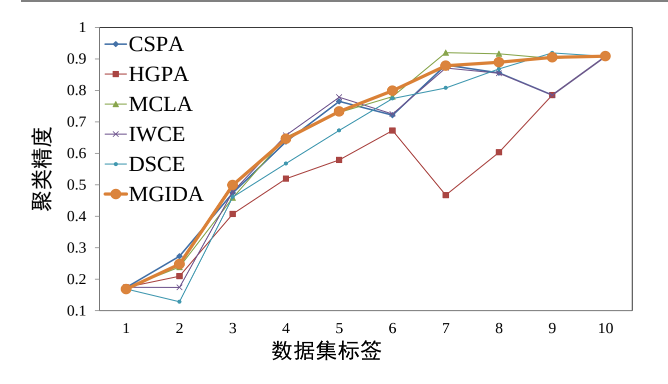


图 1 UCI 数据对比实验结果

取 100 次试验的聚类精度和方差的平均值,与基于多粒度决策不一致粗糙集的聚类融合算法（MGIDA）进行对比,使用聚类精度作为评价聚类效果的指标,得到如表 4 与图 1 的结果。由对比实验结果可以看出,MGIDA 在第 3、6、10 数据集上取得了最优的聚类精度,在第 2、4、5、7、8、9 数据集取得了次优的聚类精度或十分接近（-0.0113,第 4 个数据集，-0.0323,第 5 个数据集，-0.0023,第 7 个数据集，）,MGIDA 明显地优于 HGPA,只在第 3 个数据集上劣于 MCLA,在第 2、4、7、8、9 数据集上优于 MCLA,在其他数据集上劣于 MCLA,且与该方法精度差距不大；MGIDA 只在第 4,5 数据集上劣于 IWCE，只在第 9 数据集上劣于 DSCE.综上可知上本文的算法在每个数据集上都不是精度最差的算法。算法 2 使同一数据生成不同的弱划分本质上是在数据集上掺杂了不同程度的噪声，即通过扭曲数据的方式产生噪声,所以在这种含噪声数据的数据集上表现较好的算法具有较好的鲁棒性,本文的算法聚类精度即使不能取得最优,也可以取得或接近次优，说明本文的算法在各数据集上都有较好的表现，具有较好的鲁棒性。

算法 2 将多维数据随机地映射成一维数据，所以各算法在有的数据集上表现一般，对于分布较为复杂的数据集则表现较差，这是因为数据的扭曲使得生成的聚类成员不能很好地反映数据的真实分布所造成的.

表 5 时间效率对比表 /s

	CSPA	HGPA	MCLA	IWCE	DSCE	MGIDA
1	9.0171	0.4925	**0.3501**	396.9662	46.6548	1.044
2	73	0.7511	**0.5719**	332.7406	142.468	0.6406
3	0.5021	0.4521	0.4588	15.569	**0.0818**	0.1131
4	0.6138	0.5267	0.4719	18.7822	0.335	**0.0521**
5	0.4322	0.3861	0.414	1.8168	**0.0434**	0.0666
6	0.4252	0.4166	0.4299	1.7214	**0.0336**	0.0688
7	0.4614	0.4055	0.4211	1.8332	0.0718	**0.0695**
8	1.0599	0.4887	0.4075	4.3502	0.4854	**0.0528**
9	0.9147	0.4017	0.4132	44.731	0.3006	**0.0515**
10	25	0.6045	**0.3639**	2650	235.8422	0.4407

算法的时间效率如表 5 与图 2 所示。设每次运行时生成了h个聚类成员，每个聚类成员有K个类，每个类中有p个元素，使用多粒度粗糙集计算边界后，$|BN| = |U| - n$,则第一步计算相容度矩阵的复杂度为$(hKp)^2$，第二步 Otsu 算法的时间复杂度$|U|\log|U|$，第三步求解下近似和边界的比较次数为$(\mathrm{Kp})^2$，第四步对边界域元素重新归类的时间复杂度不大于$|U|(|BN|!)$，所以 MGIDA 的总的时间复杂度为 $O\left((\mathrm{hKp})^2) + |U|\log|U| + (\mathrm{Kp})^2 + |U|(|BN|!)\right)$,由时间复杂度可知论域大小和聚类类别数目是影响算法运行时间的最重要因素，且边界域较小时算法的时间复杂度较低。由表 5 与图 2 可得, MGIDA 在 4 个数据集上有最好的时间效率，在除第 1 数据集外的其他数据集上均取得了次优的时间效率，算法的时间复杂度较小.值得注意的是 MGIDA 在小数据集的时间效率较高，在数据集的数据量增大时表现一般，如图 2 所示。

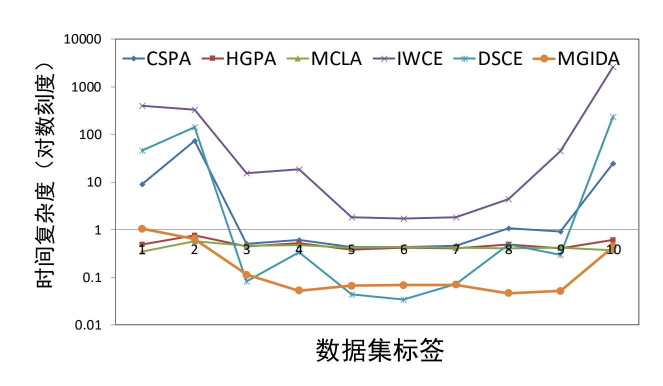


图 2 时间效率对比图

4 结束语

本文首先提出了多粒度决策不一致粗糙集模型,进而提出了一种基于多粒度粗糙集的聚类融合算法,在一个新的视角下对聚类融合算法进行了研究,进行了数据实验以与其他聚类融合算法进行对比,使用聚类精度作为指标,验证了算法的有效性。从实验结果可以看出新的聚类融合算法在部分数据集上可以取得最优,不能取得最优时,也可以取得或接近次优。本文的算法具有较好的鲁棒性.在时间效率上,本文的算法在小数据集上也具有较大优势。

K-means 算法对于非凸形分布的数据聚类效果不好,但基于本文的定理 1,对于非凸形分布的数据本文所提出的聚类融合算法受数据分布的影响较小,改进本文的算法可以应用于非凸形分布的数据,将是一个值得研究的问题.

参考文献：

[1] Qian Yuhua, Liang Jiye, Yao Yiyu, *et al.* MGRS: a multi-granulation rough set [J]. Information Sciences, 2010, 180 (6): 949-970.

[2] Lin Guoping, Liang Jiye, Qian Yuhua. An information fusion approach by combining multigranulation rough sets and evidence model [J]. Information sciences, 2015, 314: 184-199.

[3] Antoine C, Cédric W, Pierre G, *et al.* Collaborative clustering: why, when, what and how [J]. Information Fusion, 2018, 39: 81-95.

[4] 阳琳赟，王文渊. 聚类融合方法综述 [J]. 计算机应用研究, 2005, 22 (12): 8-10. (Yang Linbin, Wang Wenyuan. A survey of clustering fusion methods [J]. Application Research of Computers, 2005, 22 (12): 8-10.)

[5] Li Feijiang, Qian Yuhua, Wang Jieting, *et al.* Multigranulation information fusion: a Dempster-Shafer evidence theory-based clustering ensemble method [J]. Information Sciences, 2017, 378: 389-409.

[6] 谢岳山，樊晓平，廖志芳，等. 一种基于图论的加权聚类融合算法 [J]. 计算机应用研究, 2013, 30 (4): 1015-1016. (Xie Yueshan, Fan Xiaoping, Liao Zhifang, *et al.* A graph-based weighted clustering ensemble algorithm [j]. application research of computers, 2013, 30 (4): 1015-1016.)

[7] Fred A. Finding consistent clusters in data partitions [C]// Proc of International Workshop on Multiple Classifier Systems. Berlin: Springer, 2001: 309-318.

[8] Ayad H, Kamel M. Finding natural clusters using multi-cluster combiner based on shared nearest neighbors [C]// Proc of International Workshop on Multiple Classifier Systems. Berlin: Springer, 2003: 166-175.

[9] Faceli K, De Souto M C P, De Araújo D S A, *et al.* Multi-objective clustering ensemble for gene expression data analysis [J]. Neurocomputing, 2009, 72 (13): 2763-2774.

[10] Yi Hong, Kwong S, Wang Hanli, *et al.* Resampling-based selective clustering ensembles [J]. Pattern Recognition Letters, 2009, 30 (3): 298-305.

[11] Nguyen N, Caruana R. Consensus clusterings [C]// Proc of the 7th IEEE International Conference on Data Mining. 2007: 607-612.

[12] Otsu N. A threshold selection method from gray-level histograms [J]. IEEE Trans on Systems, Man, and Cybernetics, 1979, 9 (1): 62-66.

[13] Jain A K. Data clustering: 50 years beyond K-means [J]. Pattern Recognition Letters, 2010, 31 (8): 651-666.

[14] 阳琳赟，王路，卓晴，等．基于粗糙集理论的聚类融合加权迭代模型[J]．清华大学学报：自然科学版，2009 (8): 1106-1108. (Yang Linbin, Wang Lu, Zhuo Qing, *et al.* Weighted iteration model of clustering fusion based on rough set theory [J]. Journal of Tsinghua University: Natural Science Edition, 2009 (8): 1106-1108.)

[15] Hu Jie, Li Tianrui, Wang Hongjun, *et al.* Hierarchical cluster ensemble model based on knowledge granulation [J]. Knowledge-Based Systems, 2016, 91 (C): 179-188.

[16] Huang Dong, Wang Changdong, Lai Jianhuang. Locally Weighted Ensemble Clustering [J]. IEEE Trans on Cybernetics, 2018, 48 (5): 1460-1473.

[17] Kausar N, Abdullah A, Samir B B, *et al.* Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease [J]. Journal of Medical Imaging & Health Informatics, 2016, In-Press.

[18] Teng Geer, He Changheng, Xiao Jin, *et al.* Cluster ensemble framework based on the group method of data handling [J]. Applied Soft Computing, 2016, 43 (C): 35-46.